

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

Simon Anders¹, Davis J. McCarthy^{2,3}, Yunshen Chen^{4,5}, Michal Okoniewski⁶, Gordon K. Smyth^{4,7}, Wolfgang Huber¹ & Mark D. Robinson^{8,9}

¹*Genome Biology Unit, European Molecular Biology Laboratory, Mayerhofstrasse 1, 69117 Heidelberg, Germany*

²*Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, United Kingdom*

³*Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom*

⁴*Bioinformatics Division, Walter and Eliza Hall Institute, 1G Royal Parade, Parkville, Victoria 3052, Australia*

⁵*Department of Medical Biology, University of Melbourne, Victoria 3010, Australia*

⁶*Functional Genomics Center UNI ETH Zurich, Winterthurerstrasse 190, CH-8057, Switzerland*

⁷*Department of Mathematics and Statistics, University of Melbourne, Victoria 3010, Australia*

⁸*Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190 CH-8057 Zurich, Switzerland*

⁹*SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland*

RNA sequencing (RNA-seq) has been rapidly adopted for the profiling of transcriptomes in many areas of biology, including studies into gene regulation, development and disease. Of particular interest is the discovery of differentially expressed genes across different conditions (e.g., tissues, perturbations), while optionally adjusting for other systematic factors that affect the data collection process. There are a number of subtle yet critical aspects of these analyses, such as read counting, appropriate treatment of biological variability, quality control checks and appropriate setup of statistical modeling. Several variations have been presented in the literature, and there is a need for guidance on current best practices. This protocol presents a “state-of-the-art” computational and statistical RNA-seq differential expression analysis workflow largely based on the free open-source R language and Bioconductor software and in particular, two widely-used tools DESeq and edgeR. Hands-on time for typical small experiments

(e.g., 4-10 samples) can be <1 hour, with computation time <1 day using a standard desktop PC.

INTRODUCTION

Applications of the method. Following the wave of research using DNA microarrays, the repertoire of tools available for studying gene expression has been vastly expanded with the advent of cheap and accessible sequencing. Relative expression analyses, alternative splicing, discovery of novel transcripts and isoforms, RNA editing, allele-specific expression, the exploration of non-model organism transcriptomes, among others, make up the multitude of applications addressed by the RNA sequencing (RNA-seq) platform^{1,2}.

Several variations exist, but a typical RNA-seq experiment proceeds as follows. A sample of RNA is extracted from cells of interest. Since ribosomal RNA makes up the vast majority of cellular RNA, researchers often select poly-A messenger RNA or invoke ribosomal RNA depletion^{1,2}; existing protocols offer the capability to preserve RNA strandedness^{3,4}, and variations exist to capture small RNAs (e.g., miRNAs)⁵. From the captured RNA subpopulation, complementary DNA is synthesized and fragmented and one or both ends of these fragments are sequenced. Typically, tens of millions of sequences (“reads”) are generated, and these, across several samples, form the starting point of this protocol.

An initial and fundamental analysis goal is to identify genes that change in abundance between conditions. In the simplest case, the aim is to compare expression levels between two conditions, e.g., stimulated versus unstimulated or wild-type versus mutant. More complicated experimental designs can include additional experimental factors, potentially with multiple levels (e.g., multiple mutants, doses of a drug or time points) or may need to account for additional covariates (e.g. experimental batch or sex) or the pairing of samples (e.g., paired tumour and normal tissues from individuals).

A critical component of such an analysis is the statistical procedure used to call differentially expressed genes. This protocol covers two widely-used tools for this task: DESeq⁶ and edgeR⁷⁻¹⁰, both available as packages of the Bioconductor software development project¹¹. Applications of these methods to biology and biomedicine are many-fold. This protocol presents a workflow built from a particular set of tools, but it is modular and extensible, so alternatives that offer special features (e.g., counting by allele) or additional flexibility (e.g.,

specialized mapping strategy), can be inserted as necessary.

Development of the protocol. Figure 1 gives the overall sequence of steps, from read sequences to feature counting to the discovery of differentially expressed genes, with a concerted emphasis on quality checks throughout. After initial checks on sequence quality, reads are mapped to a reference genome with a splice-aware aligner¹²; up to this point, the Protocol is identical to many other pipelines (e.g.,¹³). From the set of mapped reads and either an annotation catalog or an assembled transcriptome, features, typically genes or transcripts, are counted and assembled into a table (rows for features and columns for samples). The pipeline is modular, and Figure 1 highlights straightforward alternative entry points to the protocol (orange boxes). The statistical methods, which are integral to the differential expression discovery task, operate on a feature count table. Before the statistical modeling, further quality checks are encouraged to ensure that the biological question can be addressed. For example, a plot of sample relations can reveal possible batch effects and can be used to understand the similarity of replicates and overall relationships between samples. After the statistical analysis of differential expression, a set of genes deemed to be differentially expressed or the corresponding statistics can be used in downstream interpretive analyses in order to confirm or generate further hypotheses.

Replication levels in designed experiments tend to be modest, often not much more than two or three. As a result, there is a need for statistical methods that perform well in small-sample situations. The low levels of replication rule out, for all practical purposes, distribution-free rank- or permutation-based methods. Thus, for small to moderate sample sizes, the strategy employed is to make formal distributional assumptions about the data observed. The advantage of parametric assumptions is the ability, through the wealth of existing statistical methodology, to make inferences about parameters of interest (i.e., changes in expression). For genome-scale count data including RNA-seq, a convenient and now well-established approximation is the negative binomial (NB) model, which represents a natural extension of the Poisson model (i.e., mixture of Gamma-distributed rates) that was used in early studies¹⁴; importantly, Poisson variation can only describe technical (i.e., sampling) variation.

The NB model has been shown to be a good fit to real data¹⁰, yet flexible enough to account for biological variability. It provides a powerful framework (e.g. via generalized linear models; GLMs) for analyzing arbitrarily complex experimental designs. NB models, as

applied to genomic count data, make the assumption that an observation, say Y_{gj} (observed number of reads for gene g and sample j), has mean μ_{gj} and variance $\mu_{gj} + \phi_g \mu_{gj}^2$, where the *dispersion* $\phi_g > 0$ represents over-dispersion relative to the Poisson distribution⁷. The mean parameters μ_{gj} depend on the sequencing depth for sample j as well as on the amount of RNA from gene g in the sample. Statistical procedures can be formulated to test for changes in expression level between experimental conditions, possibly adjusting for batch effects or other covariates, and to estimate the log-fold-changes in expression.

The dispersion ϕ_g represents the squared coefficient of variation of the true expression levels between biologically independent RNA samples under the same experimental conditions, and hence $\sqrt{\phi_g}$ is called the *biological coefficient of variation*¹⁰.

Obtaining stable estimates of the genewise dispersions is critical for reliable statistical testing. Unless the number of samples is very large, stable estimation of the dispersion requires some sort of sharing of information between genes. One can average the variability across all genes⁸, or fit an global trend to the dispersion⁶ or can seek a more general compromise between individual gene and global dispersion estimators⁷. Methods of estimating the genewise dispersion estimators have received considerable attention^{6,7,15,16}.

For the analysis of differential expression, this protocol focuses on DESeq and edgeR, which implement general differential analyses based on the NB model. These tools differ in their “look-and-feel” and estimate the dispersions somewhat differently but offer overlapping functionality (See Box 1).

BOX 1: Differences between DESeq and edgeR

The two packages described in this protocol, **DESeq** and **edgeR**, have similar strategies to perform differential analysis for count data. However, they differ in a few important areas. First, their “look-and-feel” differs. For users of the widely-used **limma** package⁴⁷ (for analysis of microarray data), the data structures and steps in **edgeR** follow analagously. The packages differ in their default normalization: **edgeR** uses the trimmed mean of M-values⁴⁸, while **DESeq** uses a “relative log expression” approach by creating a virtual library that every sample is compared against; in practice, the normalization factors are often similar. Perhaps most critical, the tools differ in the choices made to

estimate the dispersion. **edgeR** moderates feature-level dispersion estimates towards a trended mean according to the dispersion-mean relationship. In contrast, **DESeq** takes the maximum of the individual dispersion estimates and the dispersion-mean trend. In practice, this means **DESeq** is less powerful while **edgeR** is more sensitive to outliers. Recent comparison studies have highlighted that no single method dominates another across all settings^{49–51}.

Variations of the protocol. The count-based pipeline discussed here can be used in concert with other tools. For example, for species without an available well-annotated genome reference, Trinity¹⁷ or other assembly tools can be used to build a reference transcriptome; reads can then be aligned and counted, followed by the standard pipeline for differential analysis¹⁸. Similarly, to perform differential analysis on novel genes in otherwise annotated genomes, the protocol could be expanded to include merged per-sample assemblies (e.g. **cuffmerge** within the **cufflinks** package^{13,19,20}) and used as input to counting tools.

Comparison to other methods. As mentioned, the strategy taken here is to count the number of reads that fall into annotated genes and perform the statistical analysis on the table of counts to discover quantitative changes of expression levels between experimental groups. This counting approach is direct, flexible and can be used for many types of count data beyond RNA-seq, such as comparative analysis of immunoprecipitated DNA^{21–24} (e.g. ChIP-seq, MBD-seq;^{21,22}), proteomic spectral counts²⁵ and metagenomics data. Many tools exist for differential expression of counts, with slight variations of the method demonstrated in this protocol; these include, among others, **baySeq**²⁶, **BBSec**²⁷, **NOISeq**²⁸ and **QuasiSeq**²⁹.

The count-based RNA-seq analyses presented here consider the *total* output of a locus, without regard to the isoform diversity that may be present. This is of course a simplification. In certain situations, gene-level count-based methods may not recover true differential expression when some isoforms of a gene are up-regulated and others are down-regulated^{13,30}. Extensions of the gene-level count-based framework to differential exon usage are now available (e.g., **DEXSeq**³¹; discussed below). Recently, approaches have been proposed to estimate transcript-level expression and build the uncertainty of these estimates into a differential analysis at the transcript-level (e.g., **BitSeq**³²). Isoform deconvolution coupled with differential expression (e.g., **cuffdiff**^{13,19,20}) is a plausible and popular alternative, but in general, isoform-specific expression estimation remains a difficult problem, especially if sequence reads are short, if genes whose isoforms overlap substantially should be analysed, or

unless very deeply sequenced data is available. At present, isoform deconvolution methods and transcript-level differential expression methods only support two-group comparisons. In contrast, counting is straightforward, regardless of the configuration and depth of data and arbitrarily complex experiments are naturally supported through GLMs (see Box 2 for further details on feature counting). Recently, a flexible Bayesian framework for the analysis of “random” effects in the context of GLM models and RNA-seq count data was made available in the **ShrinkSeq** package¹⁵. As well, count-based methods that operate at the *exon* level, which share the same statistical framework, as well as flexible coverage-based methods have become available to address the limitations of gene-level analyses^{31,33,34}. These methods give a direct readout of differential exons, genes whose exons are used unequally, or non-parallel coverage profiles, all of which reflect a change in isoform usage.

BOX 2: Feature counting

In principle, counting reads that map to a catalog of features is straightforward. However, a few subtle decisions need to be made. For example, how should reads that fall within intronic regions (i. e., between two known exons) or beyond the annotated regions be counted? Ultimately, the answer to this question is guided by the chosen catalog that is presented to the counting software; depending on the protocol used, users should be conscious to include all features that are of interest, such as poly-adenylated RNAs, small RNAs, long intergenic non-coding RNAs and so on. For simplicity and to avoid problems with mismatching chromosome identifiers and inconsistent coordinate systems, we recommend using the curated **FASTA** files and **GTF** files from **Ensembl** or the pre-built indices packaged with **GTF** files from iGenomes³⁵, when possible.

Statistical inference based on the negative binomial distribution requires raw read counts as input. This is required to correctly model the Poisson component of the sample-to-sample variation. Therefore, it is crucial that *units of evidence* for expression are counted. No prior normalization or other transformation should be applied, including quantities such as RPKM (reads per kilobase model), FPKM (fragments per kilobase model) or otherwise depth-adjusted read counts. Both **DESeq** and **edgeR** internally keep the raw counts and normalization factors separate, as this full information is needed to correctly model the data. Notably, recent methods to normalize RNA-seq data for sample-specific G+C content effects employ offsets that are presented to the GLM, while

maintaining counts on their original scale^{36,37}.

Paired-end reads each represent a single fragment of sequenced DNA, yet (at least) two entries for the fragment will appear in the corresponding **BAM** files. Some simplistic early methods that operated on **BAM** files considered these as separate entries, which led to overcounting and would ultimately overstate the significance of differential expression.

Typically, there will be reads that cannot be uniquely assigned to a gene, either because the read was aligned to multiple locations (multi-reads) or the read's position is annotated as part of several overlapping features. For the purpose of calling differential expression, such reads should be discarded. Otherwise, genuine differential expression of one gene might cause another gene to appear differentially expressed, erroneously, if reads from the first gene are counted for the second due to assignment ambiguity. In this Protocol, we employ the tool `htseq-count` of the Python package HTSeq³⁸ using the default *union* counting mode; more details can be found at <http://www-huber.embl.de/users/anders/HTSeq/doc/count.html>. In addition, Bioconductor now offers various facilities for feature counting, including `easyRNASeq` in the `easyRNASeq` package³⁹, `summarizeOverlaps` function in the `GenomicRanges`⁴⁰ package and `qCount` in the `QuasR`⁴¹ package.

Experimental design. Some of the early RNA-seq studies were performed without biological replication. If the purpose of the experiment is to make a general statement about a biological condition of interest (in statistical parlance, a population), for example, the effect of treating a certain cell line with a particular drug, then an experiment without replication is insufficient. Rapid developments in sequencing reduce technical variation but cannot possibly eliminate biological variability⁴². Technical replicates are suited to studying properties of the RNA-seq platform¹², but they do not inform about the inherent biological variability in the system or the reproducibility of the biological result, for instance, its robustness to slight variations in cell density, passage number, drug concentration or media composition. In other words, experiments without biological replication are suited to make a statement regarding one particular sample that existed on one particular day in one particular laboratory, but not whether anybody could reproduce this result. When no replicates are available, experienced analysts may still proceed, using one of the following options: i) a descriptive analysis with no formal hypothesis testing; ii) selecting a dispersion value based on past experience; iii)

using housekeeping genes to estimate variability over all samples in the experiment.

In this context, it is helpful to remember the distinction between *designed experiments* in which a well-characterized system (e. g., a cell line or a laboratory mouse strain) undergoes a fully controlled experimental procedure with minimal unintended variation; and *observational studies*, in which samples are often those of convenience (e. g., patients arriving at a clinic) and have been subject to many uncontrolled environmental and genetic factors. Replication levels of two or three are often a practicable compromise between cost and benefit for designed experiments, whereas for observational studies typically much larger group sizes (dozens or hundreds) are needed to reliably detect biologically meaningful results.

In many cases, data are collected over time. In this situation, researchers should be mindful of factors that may unintentionally confound their result (e. g., batch effects), such as changes in reagent chemistry or software versions used to process their data⁴³. Users should make a concerted effort to: i) reduce confounding effects through experimental design (e. g., randomization, blocking⁴⁴); ii) keep track of versions, conditions (e. g., operators) of every sample, in the hope that these factors (or, surrogates of them) can be differentiated from biological factor(s) of interest in the downstream statistical modeling. In addition, there are emerging tools available that can discover and help eliminate unwanted variation^{45,46}, although these are relatively untested for RNA-seq data at present.

Complementary analyses. The focus of this protocol is gene-level differential expression analysis. However, many biologists are interested in analyses beyond that scope, and many possibilities now exist, in several cases as extensions of the count-based framework discussed here. Here, the full details of such analyses are not covered, and only a sketch of some promising approaches is made. First, an obvious extension to gene-level counting is exon-level counting, given a catalog of transcripts. Reads can be assigned to the exons that they aligned to, and these assignments be counted. Reads spanning exon-exon junctions can be counted at the junction level. The **DEXSeq** package uses a GLM that tests whether particular exons in a gene are preferentially used in a condition, over and above changes in gene-level expression. In **edgeR**, a similar strategy is taken, except that testing is done at the gene-level, effectively asking whether the exons are used proportionally across experiment conditions, in the context of biological variation.

Software implementation. There are advantages to using a small number of software plat-

forms for such a workflow, and these include simplified maintenance, training and portability. In principle, it is possible to do all computational steps in R and **Bioconductor**; however, for a few of the steps, the most mature and widely-used tools are outside **Bioconductor**. Here, R and **Bioconductor** are adopted to tie together the workflow and provide data structures, and their unique strengths in workflow components are leveraged, including statistical algorithms, visualization and computation with annotation databases. Another major advantage of an R-based system, in terms of achieving best practices in genomic data analysis, is the opportunity for an interactive analysis whereby spot checks are made throughout the pipeline to guide the analyst. In addition, a wealth of tools is available for exploring, visualizing and cross-referencing genomic data. Although not used here directly, additional features of **Bioconductor** are readily available that will often be important for scientific projects that involve an RNA-seq analysis, including access to many different file formats, range-based computations, annotation resources, manipulation of sequence data and visualisation.

In what follows, all Unix commands run at the command line appear as:

```
my_unix_command
```

whereas R functions in the text appear as **myFunction**, and (typed) R input commands and output appear as blue and orange, respectively:

```
> x = 1:10  
> median(x)
```

```
[1] 5.5
```

Note that in R, the operators = and <- can both be used for variable assignment (i. e., **z** = 5 and **z** <- 5 produce the same result, a new variable **z** with a numeric value). In this Protocol, we use the = notation; in other places, users may also see the <- notation.

File formats are denoted as PDF (i. e., for Portable Document Format).

Scope of this protocol. The aim of this Protocol is to provide a concise workflow for a standard analysis, in a complete and easily accessible format, for new users to the field or to R. We describe a specific, but very common analysis task, namely the analysis of an RNA-Seq experiment comparing two groups of samples that differ in their experimental treatment, and also cover one common complication, namely the need to account for a *blocking* factor.

In practice, users will often need to adapt this pipeline to account for the circumstances of their experiment. Especially, more complicated experimental designs will require further considerations not covered here. Therefore, we emphasize that this Protocol is not meant to replace the existing user guides, vignettes and online documentation for the packages and functions described. These provide a large body of information that is helpful to tackle tasks that go beyond the single standard workflow presented here.

In particular, `edgeR` and `DESeq` have extensive users guides, downloadable from <http://www.bioconductor.org>, that cover a wide range of relevant topics. Please consult these comprehensive resources for further details. Another rich resource for answers to commonly asked questions is the Bioconductor mailing list as well as online resources such as seqanswers.com, stackoverflow.com and biostars.org.

MATERIALS

* Equipment

Operating system: This protocol assumes users have a Unix-like operating system, i. e., Linux or MacOS X, with a bash shell or similar. All commands given here are meant to be run in a terminal window. While it is possible to follow this protocol with a Microsoft Windows machine (e. g., using Unix-like Cygwin⁵²), the additional steps required are not discussed here.

Software: Users will need the following software:

- an aligner to map short reads to a genome that is able to deal with reads that straddle introns¹². The aligner `tophat`^{19,53} is illustrated here, but others, such as `GSNAP`⁵⁴, `SpliceMap`⁵⁵ or `Subread`⁵⁶ can be used.
- optionally, a tool to visualize alignment files, such as the `Integrated Genome Viewer`

(IGV)⁵⁷, or Savant^{58,59}. IGV is a Java tool with “web start” (downloadable from <http://www.broadinstitute.org/software/igv/download>), i.e., it can be started from a web browser and needs no explicit installation at the operating system level, provided a Java Runtime Environment is available.

- the R statistical computing environment⁶⁰
- a number of Bioconductor¹¹ packages, specifically ShortRead⁶¹, DESeq⁶ and edgeR^{9,10}, and possibly GenomicRanges, GenomicFeatures and org.Dm.eg.db, as well as their dependencies.
- the samtools program⁶² (for manipulation of SAM and BAM formatted files).
- the HTSeq package³⁸ (for counting of mapped reads).
- optionally, if users wish to work with data from the Short Read Archive, the SRA Toolkit, available from <http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software>.

For many of these software packages, new features and optimizations are constantly developed and released, so it is highly recommended to use the most recent stable version as well as reading the (corresponding) documentation for the version used, since recommendations can change over time. The package versions used in the production of this article are given at the end of the protocol.

In general, the starting point is a collection of FASTQ files, the commonly used format for reads from Illumina sequencing machines. The modifications necessary for mapping reads from other platforms are not discussed here.

Example data: The data set published by Brooks et al.⁶³ is used here to demonstrate the workflow. This data set consists of seven RNA-seq samples, each a cell culture of *Drosophila melanogaster* S2 cells. Three samples were treated with siRNA targeting the splicing factor *pasilla* (CG1844) (“Knockdown”) and four samples are untreated (“Control”). Our aim is to identify genes that change in expression between Knockdown and Control.

Brooks et al.⁶³ have sequenced some of their libraries in single-end and others in paired-end mode. This allows us to demonstrate two variants of the workflow: If we ignore the differences in library type, the samples only differ by their experimental condition, knockdown or control, and the analysis is a simple comparison between two sample groups. We refer to this setting as an experiment with a *simple design*. If we want to account for library type as a blocking

factor, our samples differ in more than one aspect, i.e., we have a *complex design*. To deal with the latter, we use **edgeR** and **DESeq**'s functions to fit generalized linear models (GLMs).

* Equipment setup

Install bowtie, tophat and samtools

Download and install **samtools** from <http://samtools.sourceforge.net>.

bowtie and **tophat** have binary versions available for Linux and Mac OS X platforms. These can be downloaded from <http://bowtie-bio.sourceforge.net/index.shtml> and <http://tophat.cbcb.umd.edu>. Consult the documentation on those sites for further information if necessary. Here, **bowtie** or **bowtie2** can be used.

Install R and required Bioconductor packages

Download the **latest** version of R (at time of writing, R version 3.0.0) from <http://cran.r-project.org> and install it. Consult the **R Installation and Administration** manual if necessary.

To install **Bioconductor** packages, start R by issuing the command **R** in a terminal window and type:

```
> source( "http://www.bioconductor.org/biocLite.R" )
> biocLite("BiocUpgrade")
> biocLite( c("ShortRead", "DESeq", "edgeR") )
```

This retrieves an automatic installation tool (**biocLite**) and installs the version-matched packages. In addition, the installation tool will automatically download and install all other packages that are prerequisite. Versions of **Bioconductor** packages are matched to versions of R. Hence, to use current versions of **Bioconductor** packages, it is necessary to use a current version of R. Note that R and **Bioconductor**, at all times, maintain a stable *release* version

and a *development* version. Unless a special need exists for a particular new functionality, users should use the release version.

Download the example data

Note: This step is only required if data originate from the Short Read Archive (SRA).

Brooks et al.⁶³ deposited their data in the Short Read Archive (SRA) of the NCBI's Gene Expression Omnibus (GEO)⁶⁴ under accession number GSE18508 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18508>), and a subset of this data set will be used here to illustrate the pipeline. Specifically, SRA files corresponding to the 4 “Untreated” (Control) and 3 “CG8144_RNAi” (Knockdown) samples need to be downloaded.

For downloading SRA repository data, an automated process may be desirable. For example, from <http://www.ncbi.nlm.nih.gov/sra?term=SRP001537> (the entire experiment corresponding to GEO accession GSE18508), users can download a table of the metadata into a comma-separated tabular file “SraRunInfo.csv”. To do this, click on “Send to:” (top right corner), select “File”, select format “RunInfo” and click on “Create File”.

This CSV file “SraRunInfo.csv” is read into R, and the subset of samples that we are interested in, corresponding to the 22 SRA files shown in Figure 2, are selected (using R's string matching function `grep`) by:

```
> sri = read.csv("SraRunInfo.csv", stringsAsFactors=FALSE)
> keep = grep("CG8144|Untreated-", sri$LibraryName)
> sri = sri[keep,]
```

The “SraRunInfo.csv” file is made available in Supplementary File 1, which contains an archive of various files used in this protocol.

The following R commands automate the download of the 22 SRA files to the current working directory (the functions `getwd` and `setwd` can be used to retrieve and set the working directory, respectively):

```
> fs = basename(sri$download_path)
> for(i in 1:nrow(sri))
  download.file(sri$download_path[i], fs[i])
```

The R-based download of files described above is just one possibility of capturing several files in a semi-automatic fashion. Users can alternatively use the batch tools `wget` (Unix/Linux) or `curl` (Mac OS X), or download using a web browser. The (truncated) verbose output of the above R download commands looks as follows:

```
trying URL 'ftp://ftp-private.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR03
ftp data connection made, file length 415554366 bytes
opened URL
=====
downloaded 396.3 Mb

trying URL 'ftp://ftp-private.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR03
ftp data connection made, file length 409390212 bytes
opened URL
=====
downloaded 390.4 Mb
[... truncated ...]
```

Convert SRA to **FASTQ** format

Typically, Illumina read data from a sequencing facility will come in (compressed) **FASTQ** format. The SRA, however, uses its own, compressed, **SRA** format. To convert the example data downloaded in the previous step to **FASTQ**, use the `fastq-dump` command from the SRA Toolkit on each SRA file. R can be used to construct the required shell commands, starting from the “SraRunInfo.csv” metadata table, as follows:

```
> stopifnot( all(file.exists(fs)) ) # assure FTP download was successful
> for(f in fs) {
```

```

cmd = paste("fastq-dump --split-3", f)
cat(cmd, "\n")
system(cmd) # invoke command
}

```

Note: Users may choose to type the 22 `fastq-dump` commands **manually** into the Unix shell rather than using `R` to construct them.

It is not absolutely necessary to use `cat` to print out the current command, but it serves the purpose of knowing what is currently running in the shell:

```

fastq-dump --split-3 SRR031714.sra
Written 5327425 spots for SRR031714.sra
Written 5327425 spots total
fastq-dump --split-3 SRR031715.sra
Written 5248396 spots for SRR031715.sra
Written 5248396 spots total
[... truncated ...]

```

Be sure to use the `--split-3` option, which splits mate-pair reads into separate files. After this command, single and paired-end data will produce one or two FASTQ files, respectively. For paired-end data, the file names will be suffixed `_1.FASTQ` and `_2.FASTQ`; otherwise, a single file with extension `.FASTQ` will be produced.

Download the reference genome

Download reference genome sequence for the organism under study in FASTA format. Some useful resources, among others, include:

- the general Ensembl FTP server (<http://www.ensembl.org/info/data/ftp/index.html>)
- the Ensembl plants FTP server (<http://plants.ensembl.org/info/data/ftp/index.html>)

- the Ensembl metazoa FTP server (<http://metazoa.ensembl.org/info/data/ftp/index.html>)
- the UCSC current genomes FTP server (<ftp://hgdownload.cse.ucsc.edu/goldenPath/currentGenomes/>)

For **Ensembl**, choose the “FASTA (DNA)” link instead of “FASTA (cDNA)”, since alignments to the genome, not the transcriptome, are desired. For *Drosophila melanogaster*, the file labeled “toplevel” combines all chromosomes. Do not use the “repeat-masked” files (indicated by “rm” in the file name), since handling repeat regions should be left to the alignment algorithm.

The *Drosophila* reference genome can be downloaded from **Ensembl** and uncompressed using the following commands:

```
wget ftp://ftp.ensembl.org/pub/release-70/fasta/drosophila_melanogaster/\
dna/Drosophila_melanogaster.BDGP5.70.dna.toplevel.fa.gz
```

```
gunzip Drosophila_melanogaster.BDGP5.70.dna.toplevel.fa.gz
```

For genomes provided by UCSC, users would select their genome of interest, proceed to the “bigZips” directory and download the “chromFa.tar.gz”; as above, this could be done using the **wget** command. Note that indices for many commonly used reference genomes can be downloaded directly from <http://tophat.cbcb.umd.edu/igenomes.html> ³⁵.

Get gene model annotations

Download a GTF file with gene models for the organism of interest. For species covered by **Ensembl**, the Ensembl FTP site mentioned above contains links to such files.

The gene model annotation for *Drosophila melanogaster* can be downloaded and uncompressed using:

```
wget ftp://ftp.ensembl.org/pub/release-70/gtf/drosophila_melanogaster/\
Drosophila_melanogaster.BDGP5.70.gtf.gz
```

```
gunzip Drosophila_melanogaster.BDGP5.70.gtf.gz
```

Critical: Make sure that the gene annotation uses the same coordinate system as the reference FASTA file. Here, both files use BDGP5 (i.e., release 5 of the assembly provided by the Berkeley Drosophila Genome Project), as is apparent from the file names. To be on the safe side here, we recommend to always download the FASTA reference sequence and the GTF annotation data from the same resource provider.

As an alternative, the UCSC Table Browser (<http://genome.ucsc.edu/cgi-bin/hgTables>) can be used to generate GTF files based on a selected annotation (e.g., RefSeq genes). However, at the time of writing GTF files obtained from the UCSC Table Browser do not contain correct gene IDs, which causes problems with downstream tools such as `htseq-count`, unless corrected manually.

Build the reference index

Before reads can be aligned, the reference FASTA files need to be preprocessed into an *index* that allows the aligner easy access. To build a `bowtie2`-specific index from the FASTA file mentioned above, use the command:

```
bowtie2-build -f Drosophila_melanogaster.BDGP5.70.dna.toplevel.fa Dme1_BDGP5_70
```

A set of EBWT (or BT2 for `bowtie2`) files will be produced, with names starting with `Dme1_BDGP5_70` as specified above. This procedure needs to be run only once for each reference genome used. As mentioned, pre-built indices for many commonly-used genomes are available from <http://tophat.cbcb.umd.edu/igenomes.html> ³⁵.

PROCEDURE

Step 1 *Assess sequence quality control with **ShortRead***⁶¹

TIMING:~2 hours

[See also: *TROUBLESHOOTING*]

[See also: *ANTICIPATED RESULTS*]

At the R prompt (you may first need to use `setwd` to change to the directory where the FASTQ files are situated), type the commands:

```
> library("ShortRead")
> fqQC = qa(dirPath=".", pattern=".fastq$", type="fastq")
> report(fqQC, type="html", dest="fastqQAreport")
```

Then, use a web browser to inspect the generated HTML file (here, stored in the “fastqQAreport” directory) with the quality-assessment report.

Step 2 *Collect metadata of experimental design*

[See also: *ANTICIPATED RESULTS*]

Create a table of metadata called `samples`. This step needs to be adapted for each data set, and many users may find a spreadsheet program like `Excel` useful for this step, from which data can be imported into the table `samples` by the `read.csv` function. As an alternative, for our example data, we chose to construct the `samples` table programmatically from the table of `SRA` files. In a first step, collapse the initial table `sri` to one row per sample:

```

> sri$LibraryName = gsub("S2_DRSC_", "", sri$LibraryName) # trim label
> samples = unique(sri[, c("LibraryName", "LibraryLayout")])
> for(i in seq_len(nrow(samples))) {
  rw = (sri$LibraryName == samples$LibraryName[i])
  if(samples$LibraryLayout[i] == "PAIRED") {
    samples$fastq1[i] = paste0(sri$Run[rw], "_1.fastq", collapse="")
    samples$fastq2[i] = paste0(sri$Run[rw], "_2.fastq", collapse="")
  } else {
    samples$fastq1[i] = paste0(sri$Run[rw], ".fastq", collapse="")
    samples$fastq2[i] = ""
  }
}

```

Add important or descriptive columns to the metadata table (here, experimental groupings are set based on the “LibraryName” column, and a label is created for plotting):

```

> samples$condition = "CTL"
> samples$condition[grep("RNAi", samples$LibraryName)] = "KD"
> samples$shortname = paste( substr(samples$condition, 1, 2),
                             substr(samples$LibraryLayout, 1, 2),
                             seq_len(nrow(samples)), sep=".")

```

Carefully inspect (and correct, if necessary) the metadata table:

```

> samples

```

	LibraryName	LibraryLayout	fastq1	fastq2
1	Untreated-3	PAIRED	SRR031714_1.fastq,...	SRR031714_2.fastq,...
2	Untreated-4	PAIRED	SRR031716_1.fastq,...	SRR031716_2.fastq,...
3	CG8144_RNAi-3	PAIRED	SRR031724_1.fastq,...	SRR031724_2.fastq,...
4	CG8144_RNAi-4	PAIRED	SRR031726_1.fastq,...	SRR031726_2.fastq,...
5	Untreated-1	SINGLE	SRR031708.fastq,...	
6	CG8144_RNAi-1	SINGLE	SRR031718.fastq,...	
7	Untreated-6	SINGLE	SRR031728.fastq,...	
	condition	shortname		

1	CTL	CT.PA.1
2	CTL	CT.PA.2
3	KD	KD.PA.3
4	KD	KD.PA.4
5	CTL	CT.SI.5
6	KD	KD.SI.6
7	CTL	CT.SI.7

Step 3 *Align the reads (using **tophat**) to reference genome*

TIMING:~45 minutes per sample

[See also: TROUBLESHOOTING]

[See also: ANTICIPATED RESULTS]

Using R string manipulation, construct the Unix commands to call **tophat**. Given the meta-data table **samples**, it is convenient to use R to create the list of shell commands, as follows:

```
> gf = "Drosophila_melanogaster.BDGP5.70.gtf"
> bowind = "Dme1_BDGP5_70"
> cmd = with(samples,
  paste("tophat -G", gf, "-p 5 -o", LibraryName, bowind,
    fastq1, fastq2))
```

```
> cmd
```

```
tophat -G Drosophila_melanogaster.BDGP5.70.gtf -p 5 -o Untreated-3 \
Dme1_BDGP5_70 SRR031714_1.fastq,SRR031715_1.fastq \
SRR031714_2.fastq,SRR031715_2.fastq
```

```
tophat -G Drosophila_melanogaster.BDGP5.70.gtf -p 5 -o Untreated-4 \
Dme1_BDGP5_70 SRR031716_1.fastq,SRR031717_1.fastq \
```

SRR031716_2.fastq,SRR031717_2.fastq

```
tophat -G Drosophila_melanogaster.BDGP5.70.gtf -p 5 -o CG8144_RNAi-3 \
Dme1_BDGP5_70 SRR031724_1.fastq,SRR031725_1.fastq \
SRR031724_2.fastq,SRR031725_2.fastq
```

```
tophat -G Drosophila_melanogaster.BDGP5.70.gtf -p 5 -o CG8144_RNAi-4 \
Dme1_BDGP5_70 SRR031726_1.fastq,SRR031727_1.fastq \
SRR031726_2.fastq,SRR031727_2.fastq
```

```
tophat -G Drosophila_melanogaster.BDGP5.70.gtf -p 5 -o Untreated-1 \
Dme1_BDGP5_70 \
SRR031708.fastq,SRR031709.fastq,SRR031710.fastq,SRR031711.fastq,SRR031712.fastq,SRR03171
```

```
tophat -G Drosophila_melanogaster.BDGP5.70.gtf -p 5 -o CG8144_RNAi-1 \
Dme1_BDGP5_70 \
SRR031718.fastq,SRR031719.fastq,SRR031720.fastq,SRR031721.fastq,SRR031722.fastq,SRR03172
```

```
tophat -G Drosophila_melanogaster.BDGP5.70.gtf -p 5 -o Untreated-6 \
Dme1_BDGP5_70 SRR031728.fastq,SRR031729.fastq
```

Run these commands (i.e. copy-and-paste) in a Unix terminal.

Note: Users can either use the R function `system` to execute these commands, cut-and-paste the commands into a separate Unix shell. In addition, users could construct the unix commands independent of R.

Step 4 Organize, sort and index the BAM files and create SAM files

TIMING:~1 hour

Organize the BAM files into a single directory, sort and index them and create SAM files, as follows. (Here, we use R string manipulation to generate commands; alternatively, users

may choose to create the shell commands manually in a text editor.)

```
> for(i in seq_len(nrow(samples))) {  
  lib = samples$LibraryName[i]  
  ob = file.path(lib, "accepted_hits.bam")  
  
  # copy file  
  cat(paste0("cp ",ob," ",lib,".bam"),"\n")  
  
  # sort by name  
  cat(paste0("samtools sort -n ",lib,".bam ",lib,"_sn"),"\n")  
  
  # convert to SAM for htseq-count  
  cat(paste0("samtools view -o ",lib,"_sn.sam ",lib,"_sn.bam"),"\n")  
  
  # sort by position  
  cat(paste0("samtools sort ",ob," ",lib,"_s"),"\n")  
  
  # for IGV  
  cat(paste0("samtools index ",lib,"_s.bam"),"\n\n")  
}
```

```
cp Untreated-3/accepted_hits.bam Untreated-3.bam  
samtools sort -n Untreated-3.bam Untreated-3_sn  
samtools view -o Untreated-3_sn.sam Untreated-3_sn.bam  
samtools sort Untreated-3/accepted_hits.bam Untreated-3_s  
samtools index Untreated-3_s.bam
```

```
cp Untreated-4/accepted_hits.bam Untreated-4.bam  
samtools sort -n Untreated-4.bam Untreated-4_sn  
samtools view -o Untreated-4_sn.sam Untreated-4_sn.bam  
samtools sort Untreated-4/accepted_hits.bam Untreated-4_s  
samtools index Untreated-4_s.bam
```

```
cp CG8144_RNAi-3/accepted_hits.bam CG8144_RNAi-3.bam
```

```
samtools sort -n CG8144_RNAi-3.bam CG8144_RNAi-3_sn
samtools view -o CG8144_RNAi-3_sn.sam CG8144_RNAi-3_sn.bam
samtools sort CG8144_RNAi-3/accepted_hits.bam CG8144_RNAi-3_s
samtools index CG8144_RNAi-3_s.bam
```

```
cp CG8144_RNAi-4/accepted_hits.bam CG8144_RNAi-4.bam
samtools sort -n CG8144_RNAi-4.bam CG8144_RNAi-4_sn
samtools view -o CG8144_RNAi-4_sn.sam CG8144_RNAi-4_sn.bam
samtools sort CG8144_RNAi-4/accepted_hits.bam CG8144_RNAi-4_s
samtools index CG8144_RNAi-4_s.bam
```

```
cp Untreated-1/accepted_hits.bam Untreated-1.bam
samtools sort -n Untreated-1.bam Untreated-1_sn
samtools view -o Untreated-1_sn.sam Untreated-1_sn.bam
samtools sort Untreated-1/accepted_hits.bam Untreated-1_s
samtools index Untreated-1_s.bam
```

```
cp CG8144_RNAi-1/accepted_hits.bam CG8144_RNAi-1.bam
samtools sort -n CG8144_RNAi-1.bam CG8144_RNAi-1_sn
samtools view -o CG8144_RNAi-1_sn.sam CG8144_RNAi-1_sn.bam
samtools sort CG8144_RNAi-1/accepted_hits.bam CG8144_RNAi-1_s
samtools index CG8144_RNAi-1_s.bam
```

```
cp Untreated-6/accepted_hits.bam Untreated-6.bam
samtools sort -n Untreated-6.bam Untreated-6_sn
samtools view -o Untreated-6_sn.sam Untreated-6_sn.bam
samtools sort Untreated-6/accepted_hits.bam Untreated-6_s
samtools index Untreated-6_s.bam
```

Run these commands in a Unix terminal.

Step 5 *Inspect alignments with IGV*

Start **IGV**, select the correct genome (here, *D. melanogaster* (*dm3*)) and load the **BAM** files (here, those with the `_s` in the filename) and the **GTF** file. Zoom in on an expressed transcript until individual reads are shown and check whether the reads align at and across exon-exon junctions, as expected given the annotation (See example in Figure 3). If any positive and negative controls are known for the system under study, direct the IGV browser to these regions to confirm what is previously known.

Step 6 *Count reads using htseq-count*

TIMING:~3 hours

[See also: **TROUBLESHOOTING**]

Add the names of the **COUNT** files to the metadata table and call the command line tool **HTSeq**³⁸. Again, **R** can be used to construct the commands.

```
> samples$countf = paste(samples$LibraryName, "count", sep=".")

> gf = "Drosophila_melanogaster.BDGP5.70.gtf"
> cmd = paste0("htseq-count -s no -a 10 ", samples$LibraryName, "_sn.sam ",
               gf, " > ", samples$countf)

> cmd
```

```
htseq-count -s no -a 10 Untreated-3_sn.sam \
Drosophila_melanogaster.BDGP5.70.gtf > Untreated-3.count
```

```
htseq-count -s no -a 10 Untreated-4_sn.sam \  
Drosophila_melanogaster.BDGP5.70.gtf > Untreated-4.count
```

```
htseq-count -s no -a 10 CG8144_RNAi-3_sn.sam \  
Drosophila_melanogaster.BDGP5.70.gtf > CG8144_RNAi-3.count
```

```
htseq-count -s no -a 10 CG8144_RNAi-4_sn.sam \  
Drosophila_melanogaster.BDGP5.70.gtf > CG8144_RNAi-4.count
```

```
htseq-count -s no -a 10 Untreated-1_sn.sam \  
Drosophila_melanogaster.BDGP5.70.gtf > Untreated-1.count
```

```
htseq-count -s no -a 10 CG8144_RNAi-1_sn.sam \  
Drosophila_melanogaster.BDGP5.70.gtf > CG8144_RNAi-1.count
```

```
htseq-count -s no -a 10 Untreated-6_sn.sam \  
Drosophila_melanogaster.BDGP5.70.gtf > Untreated-6.count
```

Run these commands in a Unix terminal.

Step 7 *For differential expression analysis with **edgeR**, follow Box 3.*

Step 8 *For differential expression analysis with **DESeq**, follow Box 4.*

BOX 3: Differential count analysis with edgeR

Step 1 edgeR *Create container for count data and filter features*

Load the `edgeR` package and use the utility function, `readDGE`, to read in the COUNT files created from `htseq-count`:

```
> library("edgeR")
> counts = readDGE(samples$countf)$counts
```

In `edgeR`, we recommend removing features without at least 1 read per million in n of the samples, where n is the size of the smallest group of replicates (here, $n = 3$ for the Knockdown group). Filter these as well as non-informative (e. g., non-aligned) features using a command like:

```
> noint = rownames(counts) %in%
      c("no_feature", "ambiguous", "too_low_aQual",
        "not_aligned", "alignment_not_unique")
> cpms = cpm(counts)
> keep = rowSums(cpms>1)>=3 & !noint
> counts = counts[keep,]
```

Visualize and inspect the count table using:

```
> colnames(counts) = samples$shortname
> head( counts[,order(samples$condition)], 5 )
```

	CT.PA.1	CT.PA.2	CT.SI.5	CT.SI.7	KD.PA.3	KD.PA.4	KD.SI.6
FBgn0000008	76	71	137	82	87	68	115
FBgn0000017	3498	3087	7014	3926	3029	3264	4322
FBgn0000018	240	306	613	485	288	307	528
FBgn0000032	611	672	1479	1351	694	757	1361
FBgn0000042	40048	49144	97565	99372	70574	72850	95760

Create a `DGEList` object (`edgeR`'s container for RNA-seq count data), as follows:

```
> d = DGEList(counts=counts, group=samples$condition)
```

Step 2 `edgeR` *Estimate normalization factors*

Estimate normalization factors using:

```
> d = calcNormFactors(d)
> d$samples
```

Step 3 `edgeR` *Inspect sample relations*

Use the `plotMDS` function to create a count-specific multidimensional scaling plot (shown in Figure 4A):

```
> plotMDS(d, labels=samples$shortname,
          col=c("darkgreen", "blue")[factor(samples$condition)])
```

Step 4 `edgeR` *Estimate dispersion and conduct statistical tests according to A if a simple two-group comparison, or B if a complex design.*

[See also: TROUBLESHOOTING]

A. Perform statistical calculations for a simple two-group comparison

(i) Estimate dispersions (“classic” `edgeR`)

[See also: TROUBLESHOOTING]

For simple designs, estimate tagwise dispersion estimates using:

```
> d = estimateCommonDisp(d)
> d = estimateTagwiseDisp(d)
```

Create a visual representation of the mean-variance relationship using `plotMeanVar` (shown in Figure 6A), as follows:

```
> plotMeanVar(d, show.tagwise.vars=TRUE, NBline=TRUE)
```

and `plotBCV` (Figure 6B), as follows:

```
> plotBCV(d)
```

(ii) Test for differential expression (“classic” `edgeR`)

[See also: TROUBLESHOOTING]

For a simple two-group design, perform an exact test for the difference in expression between the two conditions:

```
> de = exactTest(d, pair=c("CTL","KD"))
```

B. Perform statistical calculations for a complex design

(i) Estimate dispersions (“GLM” `edgeR`)

[See also: TROUBLESHOOTING]

For more complex designs, create a design matrix to specify the factors that are expected to affect expression levels:

```
> design = model.matrix( ~ LibraryLayout + condition, samples)
> design
```

```
      (Intercept) LibraryLayoutSINGLE conditionKD
1             1             0             0
2             1             0             0
3             1             0             1
4             1             0             1
5             1             1             0
6             1             1             1
7             1             1             0
attr(,"assign")
[1] 0 1 2
attr(,"contrasts")
attr(,"contrasts")$LibraryLayout
[1] "contr.treatment"

attr(,"contrasts")$condition
[1] "contr.treatment"
```

Estimate dispersion values, relative to the design matrix, using the Cox-Reid (CR) adjusted likelihood^{10,65}, as follows:

```
> d2 = estimateGLMTrendedDisp(d, design)
> d2 = estimateGLMTagwiseDisp(d2, design)
```

(ii) Test for differential expression (“GLM” edgeR)

[See also: TROUBLESHOOTING]

Given the design matrix and dispersion estimates, fit the GLM to the data:

```
> f = glmFit(d2, design)
```

Perform a likelihood ratio test, specifying the difference of interest (here, Knockdown versus Control, corresponding to the 3rd column of the design matrix):

```
> lrt = glmLRT(f, coef=3)
```

Step 5 edgeR *Inspect the results in graphical and tabular format*

Use the `topTags` function to present a tabular summary of the differential expression statistics (Note: `topTags` operates on the output of either `exactTest` or `glmLRT`, while only the latter is shown here):

```
> tt = topTags(lrt, n=nrow(d))
> head(tt$table)
```

	logFC	logCPM	LR	PValue	FDR
FBgn0039155	-4.61	5.87	902	3.94e-198	2.84e-194
FBgn0025111	2.87	6.86	641	2.07e-141	7.45e-138
FBgn0039827	-4.05	4.40	457	2.08e-101	4.98e-98
FBgn0035085	-2.58	5.59	408	9.56e-91	1.72e-87
FBgn0000071	2.65	4.73	365	2.52e-81	3.63e-78
FBgn0003360	-3.12	8.42	359	4.43e-80	5.31e-77

Inspect the depth-adjusted reads per million some of the top differentially expressed genes:

```
> nc = cpm(d, normalized.lib.sizes=TRUE)
> rn = rownames(tt$table)
> head(nc[rn,order(samples$condition)],5)
```

	CT.PA.1	CT.PA.2	CT.SI.5	CT.SI.7	KD.PA.3	KD.PA.4	KD.SI.6
FBgn0039155	91.07	98.0	100.75	106.78	3.73	4.96	3.52
FBgn0025111	34.24	31.6	26.64	28.46	247.43	254.28	188.39
FBgn0039827	39.40	36.7	30.09	34.47	1.66	2.77	2.01
FBgn0035085	78.06	81.4	63.59	74.08	13.49	14.13	10.99
FBgn0000071	9.08	9.2	7.48	5.85	52.08	55.93	45.65

Create a graphical summary, such as an M (log-fold-change) versus A (log-average-expression) plot⁶⁶, here showing the genes selected as differentially expressed (with a 5% false discovery rate; see Figure 5A):

```
> deg = rn[tt$table$FDR < .05]
> plotSmear(d, de.tags=deg)
```

Step 6 edgeR *Create persistent storage of results*

Save the result table as a CSV (comma-separated values) file (alternative formats are possible) as follows:

```
> write.csv( tt$table, file="toptags_edgeR.csv" )
```

BOX 4: Differential count analysis with DESeq

Step 1 DESeq *Create container for count data*

Create a `data.frame` with the required metadata, i. e., the names of the count files and experimental conditions. Here, we derive it from the `samples` table created earlier.

```
> samplesDESeq = with(samples, data.frame(
  shortname      = I(shortname),
  countf         = I(countf),
  condition      = condition,
  LibraryLayout  = LibraryLayout))
```

Load the DESeq package and create a `CountDataSet` object (DESeq's container for RNA-seq data) from the count tables and corresponding metadata:

```
> library("DESeq")
> cds = newCountDataSetFromHTSeqCount( samplesDESeq )
```

Step 2 DESeq *Estimate normalization factors*

Estimate normalization factors using:

```
> cds = estimateSizeFactors( cds )
```

Inspect the size factors using:

```
> sizeFactors( cds )
```

```
CT.PA.1 CT.PA.2 KD.PA.3 KD.PA.4 CT.SI.5 KD.SI.6 CT.SI.7
 0.699   0.811   0.822   0.894  1.543   1.372   1.104
```

Step 3 DESeq *Inspect sample relations*

To inspect sample relationships, invoke a variance stabilizing transformation and inspect a principal component analysis (PCA) plot (shown in Figure 4B):

```
> cdsB = estimateDispersions(cds, method="blind")
> vsd = varianceStabilizingTransformation(cdsB)
> p = plotPCA(vsd, intgroup=c("condition", "LibraryLayout"))
```

Step 4 DESeq *Estimate dispersions, conduct statistical tests and inspect results according to A if a simple two-group comparison, or B if a complex design.*

[See also: TROUBLESHOOTING]

A. Perform statistical calculation for a simple two-group comparison

(i) Estimate dispersions

For simple designs, use `estimateDispersions` to calculate dispersion estimates:

```
> cds = estimateDispersions(cds)
```

Inspect the estimated dispersions using the `plotDispEsts` function (shown in Figure 6C), as follows:

```
> plotDispEsts(cds)
```

(ii) Test for differential expression

[See also: TROUBLESHOOTING]

Perform the test for differential expression, using `nbinomTest`, as follows:

```
> res = nbinomTest(cds,"CTL","KD")
```

(iii) Inspect the results in graphical and tabular format

Inspect the result table (ordered by adjusted P-value) using:

```
> head(res[order(res$padj),])
```

	id	baseMean	baseMeanA	baseMeanB	log2FoldChange	pval
9499	FBgn0039155	684	1161	47.6	-4.61	3.05e-152
2310	FBgn0025111	1355	354	2689.5	2.92	5.37e-107
9967	FBgn0039827	246	412	25.1	-4.04	1.95e-82
595	FBgn0003360	4013	6488	713.3	-3.19	5.95e-76
2561	FBgn0026562	40660	62008	12195.3	-2.35	3.96e-69
12551	FBgn0058469	1036	456	1808.6	1.99	1.25e-64
	padj					
9499	3.88e-148					
2310	3.42e-103					
9967	8.28e-79					
595	1.90e-72					
2561	1.01e-65					
12551	2.65e-61					

Count the number of genes with significant differential expression at false discovery rate (FDR) of 10%:

```
> table( res$padj < 0.1 )
```

FALSE	TRUE
11861	885

Given the table of differential expression results, use `plotMA` to display differential expression (log-fold-changes) versus expression strength (log-average-read-count), as follows (see Figure 5B):

```
> plotMA(res)
```

B. Perform statistical calculations for a complex design

(i) Estimate dispersions (complex design) *[See also: TROUBLESHOOTING]*

For complex designs, calculate the CR adjusted profile likelihood⁶⁵ dispersion estimates, developed by McCarthy et al.¹⁰, according to:

```
> cds = estimateDispersions( cds, method = "pooled-CR",  
                             modelFormula = count ~ LibraryLayout + condition )
```

(ii) Test for differential expression

[See also: TROUBLESHOOTING]

Test for differential expression in the GLM setting by fitting both a full model and reduced model (i. e., with the factor of interest taken out):

```
> fit1 = fitNbinomGLMs( cds, count ~ LibraryLayout + condition )  
> fit0 = fitNbinomGLMs( cds, count ~ LibraryLayout )
```

Using the two fitted models, compute likelihood ratio statistics and associated P-values, as follows:

```
> pval = nbinomGLMTest( fit1, fit0 )
```

(iii) Inspect the results in graphical and tabular format

Adjust the reported p values for multiple testing.

```
> padj = p.adjust( pval, method="BH" )
```

Assemble a result table from the fit of the full model and the raw and adjusted p values and print the first few lines of the table after sorting by p value, in order to inspect the top hits.

```
> res = cbind( fit1, pval = pval, padj = padj )
> head(res[order(res$padj),])
```

	(Intercept)	LibraryLayoutSINGLE	conditionKD	deviance	converged	pval
FBgn0000071	6.69	-0.348	2.65	1.69	TRUE	0
FBgn0001137	9.13	0.146	-1.24	3.72	TRUE	0
FBgn0001224	8.33	-0.137	1.45	4.21	TRUE	0
FBgn0001225	7.98	-0.257	1.30	2.25	TRUE	0
FBgn0001226	9.32	-0.343	1.66	4.10	TRUE	0
FBgn0002868	6.82	0.727	-2.20	1.55	TRUE	0

	padj
FBgn0000071	0
FBgn0001137	0
FBgn0001224	0
FBgn0001225	0
FBgn0001226	0
FBgn0002868	0

Count the number of hits at 10% false discovery rate.

```
> table( res$pval < 0.1 )
```

```
FALSE TRUE  
10457 2289
```

Step 5 DESeq *Create persistent storage of results*

Save the result to a CSV file.

```
> write.csv( res, file="res_DESeq.csv" )
```

Step 9 *Quality check of the differential expression analysis results*

Perform a sanity check by inspecting a histogram of unadjusted p -values (see Figure 7) for the differential expression results, as follows:

```
> hist(res$pval, breaks=100)
```

In addition, users should point their data browser (e. g., IGV) to a handful of the top differentially expressed genes, to double check that counting and differential expression statistics have been successful.

VERSIONS

The preprint of this document was produced with Sweave⁶⁷ using the following versions of R and its packages:

```
> sessionInfo()
```

R version 3.0.0 (2013-04-03)

Platform: x86_64-unknown-linux-gnu (64-bit)

locale:

```
[1] LC_CTYPE=en_CA.UTF-8      LC_NUMERIC=C
[3] LC_TIME=en_CA.UTF-8       LC_COLLATE=en_CA.UTF-8
[5] LC_MONETARY=en_CA.UTF-8   LC_MESSAGES=en_CA.UTF-8
[7] LC_PAPER=C                LC_NAME=C
[9] LC_ADDRESS=C              LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_CA.UTF-8 LC_IDENTIFICATION=C
```

attached base packages:

```
[1] parallel stats      graphics grDevices utils      datasets methods
[8] base
```

other attached packages:

```
[1] DESeq_1.12.0      lattice_0.20-15    locfit_1.5-9       Biobase_2.19.3
[5] BiocGenerics_0.6.0 edgeR_3.2.0        limma_3.16.0       cacheSweave_0.6-1
[9] stashR_0.3-5      filehash_2.2-1
```

loaded via a namespace (and not attached):

```
[1] annotate_1.38.0      AnnotationDbi_1.21.17 DBI_0.2-5
[4] digest_0.6.3        genefilter_1.42.0    geneplotter_1.38.0
[7] grid_3.0.0          IRanges_1.18.0       RColorBrewer_1.0-5
[10] RSQLite_0.11.2      splines_3.0.0        stats4_3.0.0
[13] survival_2.37-4     tools_3.0.0          XML_3.96-1.1
[16] xtable_1.7-1
```

The versions of software packages used can be captured with the following commands:

```
> system("bowtie2 --version | grep align", intern=TRUE)
```

```
[1] "/usr/local/software/bowtie2-2.1.0/bowtie2-align version 2.1.0"
```

```
> system("tophat --version", intern=TRUE)
```

```
[1] "TopHat v2.0.8"
```

```
> system("htseq-count | grep version", intern=TRUE)
```

```
[1] "framework, version 0.5.3p9."
```

```
> system("samtools 2>&1 | grep Version", intern=TRUE)
```

```
[1] "Version: 0.1.18 (r982:295)"
```

TIMING

Running this protocol on the SRA-downloaded data will take ~ 10 hours on a machine with eight cores and 8 GB of RAM; with a machine with more cores, mapping of different samples can be run simultaneously. The time is largely spent on quality checks of reads, read alignment and feature counting; computation time for the differential expression analysis is comparatively smaller.

Step 1, Sequence quality checks, ~ 2 h

Step 2, Organizing metadata: $\sim < 1$ h

Steps 3-5, Read alignment: ~ 6 h

Step 6, Feature counting: ~ 3 h

Step 7-14, Differential analysis: variable; computational time is often < 20 min

TROUBLESHOOTING

Step(s)	Problem	Possible reason	Solution
1,7	An error occurs when loading a Bioconductor package	Version mismatch	Make sure the most recent version of R is installed; reinstall packages using biocLite
3	An error occurs while mapping reads to reference genome	Wrong files made available or version mismatch	Carefully check the command submitted, the documentation for the aligner and the setup steps (e.g., building an index); check that there is no clash between bowtie and bowtie2
6	An error occurs counting features	GTF format violation	Use an Ensembl GTF format or coerce your file into a compatible format. In particular, verify that each line of type <i>exon</i> contains attributes named <i>gene_id</i> and <i>transcript_id</i> , and that their values are correct.
10-11	Errors in fitting statistical models or running statistical tests	Wrong inputs, outdated version of software	Ensure versions of R and Bioconductor packages are up to date and check the command issued; if command is correct and error persists, post a message to Bioconductor mailing list ⁶⁸ following the posting guide ⁶⁹ .

ANTICIPATED RESULTS

Multiple entry points to the protocol. As mentioned, this protocol is modular, in that users can use an alternative aligner, or a different strategy (or software package) to count features. Two notable entry points (See orange boxes in Figure 1) for the protocol include starting with either: i) a set of **SAM/BAM** files from an alternative alignment algorithm; ii) a

table of counts. With SAM/BAM files in hand, users can start at Step 6, although it is often invaluable to carry along metadata information (Step 2) and post-processing the alignment files may still be necessary (Step 4). With a count table in hand, users can start at Step 7 or Step 8, where again the metadata information (Step 2) will be needed for the statistical analysis. Supplementary File 1 is an archive containing: the intermediate COUNT files used here, a collated count table (`counts`) in CSV format, the metadata table (`samples`) in CSV format and the CSV file that was downloaded from the NCBI's Short Read Archive.

Sequencing quality checks. Step 1 results in an HTML report for all included FASTQ files. Users should inspect these and look for persistence of low quality scores, overrepresentation of adapter sequence and other potential problems. From these inspections, users may choose to remove low-quality samples, trim ends of reads (e.g., using FASTX⁷⁰) or modify alignment parameters. Note that a popular non-Bioconductor alternative for sequencing quality checks is FastQC⁷¹.

Metadata. In general, it is recommended to start from a sample metadata table that contains sample identifiers, experimental conditions, blocking factors and file names. In our example, we construct this table from the “SraRunInfo” CSV file provided by the Short Read Archive (SRA; Step 2). Users will often obtain a similar table from a local laboratory information management system (LIMS) and can adapt this strategy to their own data sets.

Mapping reads to reference genome. In Step 3, R is used to tie the pipeline together (i.e., loop through the set of samples and construct the full `tophat` command), with the hope of reducing typing and copy-and-paste errors. Many alternatives and variations are possible: users can use R to create and call the `tophat` commands, or just to create the commands (and call `tophat` independently from a Unix shell), or assemble the commands manually independent of R.

In the call to `tophat`, the option `-G` points `tophat` to a GTF file of annotation to facilitate mapping reads across exon-exon junctions (some of which can be found *de novo*), `-o` specifies the output directory, `-p` specifies the number of threads to use (this may affect run times and can vary depending on the resources available). Other parameters can be specified here, as needed; see the appropriate documentation for the tool and version you are using. The first argument, `Dmel_BDGP5_70` is the name of the index (built in advance), and the second argument is a list of all FASTQ files with reads for the sample. Note that the

FASTQ files are concatenated with commas, *without* spaces. For experiments with paired-end reads, pairs of FASTQ files are given as separate arguments and the order in both arguments must match.

`tophat` creates a directory for each sample with the mapped reads in a BAM file, called *accepted_hits.bam*. Note that BAM (Binary Alignment Map) files, and equivalently SAM (Sequence Alignment/Map; an uncompressed text version of BAM) are the *de facto* standard file for alignments. Therefore, alternative mapping tools that produce BAM/SAM files could be inserted into the Protocol at this Step.

Organizing BAM and SAM files. The set of *accepted_hits.bam* files (typically) need to be transformed before they can be used with other downstream tools. In Step 4, the `samtools` command was used to prepare variations of the mapped reads. Specifically, a sorted and indexed version of the BAM file was created, which can be used in genome browsers such as IGV; a sorted-by-name SAM file was created, which is compatible with the feature counting software of `htseq-count`. Alternative feature counting tools (e.g., in `Bioconductor`) may require different inputs.

Feature counting. In Step 6, we used `htseq-count` for feature counting. In particular, the option `-s` signifies that the data is not from a stranded protocol (this may vary by experiment) and the `-a` option specifies a minimum score for the alignment quality. The output is a COUNT file (2-columns: identifier, count) for each sample. Many alternatives exist inside and outside of `Bioconductor` to arrive at a table of counts given BAM (or SAM) files and a set of features (e.g., from a GTF file); see Box 2 for further considerations. Each cell in the count table will be an integer that indicates how many reads in the sample overlap with the respective feature. Non-informative rows, such as features that are not of interest or those that have low overall counts can be filtered. Such filtering (so long as it is independent of the test statistic) is typically beneficial for the statistical power of the subsequent differential expression analysis⁷².

“Normalization”. As different libraries will be sequenced to different depths, the count data are scaled (in the statistical model) so as to be comparable. The term *normalization* is often used for that, but it should be noted that the raw read counts are not actually altered⁴⁸. By default, `edgeR` uses the number of mapped reads (i.e., count table column sums) and estimates an additional normalization factor to account for sample-specific effects (e.g., diver-

sity)⁴⁸; these two factors are combined and used as an *offset* in the NB model. Analogously, DESeq defines a *virtual* reference sample by taking the median of each gene’s values across samples, and then computes *size factors* as the median of ratios of each sample to the reference sample. Generally, the ratios of the size factors should roughly match the ratios of the library sizes. Dividing each column of the count table by the corresponding size factor yields normalized count values, which can be scaled to give a *counts per million* interpretation (see also edgeR’s `cpm` function). From an M (log-ratio) versus A (log-expression-strength) plot, count datasets typically show a (left-facing) trombone shape, reflecting the higher variability of log-ratios at lower counts (See Figure 5). In addition, points will typically be centered around a log-ratio of 0 if the normalization factors are calculated appropriately, although this is just a general guide.

Sample relations. The quality of the sequencing reactions (Step 1) themselves are only part of the quality assessment procedure. In Step 3, a “fitness for use”⁷³ check is performed (relative to the biological question of interest) on the count data before statistical modeling. edgeR adopts a straightforward approach that compares the relationship between all pairs of samples, using a count-specific pairwise distance measure (i. e., biological coefficient of variation) and an MDS plot for visualization (Figure 4A). Analogously, DESeq performs a variance-stabilizing transformation and explores sample relationships using a PCA plot (Figure 4B). In either case, the analysis for the current data set highlights that library type (single-end or paired-end) has a systematic effect on the read counts and provides an example of a data-driven modeling decision: here, a GLM-based analysis that accounts for the (assumed linear) effect of library type jointly with the biological factor of interest (i. e., Knockdown versus Control) is recommended. In general, users should be conscious that the degree of variability between biological replicates (e. g., in an MDS or PCA plot) will ultimately impact the calling of differential expression. For example, a single outlying sample may drive increased dispersion estimates and compromise the discovery of differentially expressed features. No general prescription is available for when and whether to delete outlying samples.

Design matrix. For more complex designs (i. e., beyond two-group comparisons), users need to provide a design matrix that specifies the factors that are expected to affect expression levels. As mentioned above, GLMs can be used to analyze arbitrarily complex experiments, and the design matrix is the means by which the experimental design is described mathematically, including both biological factors of interest and other factors not of direct interest,

such as batch effects. For example, Section 4.5 of the **edgeR** User’s Guide or Section 4 of the **DESeq** vignette present worked case studies with batch effects. The design matrix is central for such more complex differential expression analyses, and users may wish to consult with a linear modeling textbook⁷⁴ or with a local statistician to make sure their design matrix is appropriately specified.

Dispersion estimation. As mentioned above, getting good estimates of the dispersion parameter is critical to the inference of differential expression. For simple designs, **edgeR** uses the quantile-adjusted conditional maximum (weighted) likelihood estimator^{7,8}, whereas **DESeq** uses a method-of-moments estimator⁶. For complex designs, the dispersion estimates are made relative to the design matrix, using the CR adjusted likelihood^{10,65}; both **DESeq** and **edgeR** use this estimator. **edgeR**’s estimates are always moderated toward a common trend, whereas **DESeq** chooses the maximum of the individual estimate and a smooth fit (dispersion versus mean) over all genes. A wide range of dispersion-mean relationships exist in RNA-seq data, as viewed by **edgeR**’s [plotBCV](#) or **DESeq**’s [plotDispEsts](#); case studies with further details are presented in both **edgeR**’s and **DESeq**’s user guides.

Differential expression analysis. **DESeq** and **edgeR** differ slightly in the format of results outputted, but each contain columns for (log) fold change, (log) counts-per-million (or mean by condition), likelihood ratio statistic (for GLM-based analyses), as well as raw and adjusted p-values. By default, P-values are adjusted for multiple testing using the Benjamini-Hochberg⁷⁵ procedure. If users enter tabular information to accompany the set of features (e.g. annotation information), **edgeR** has a facility to carry feature-level information into the results table.

Persistent storage of analysis results. In Box 3 Step 6 or Box 4 Step 5, [write.csv](#) is used to save results tables to disk, but can also be used to persistently store the count table (`d$counts`), the normalized counts or metadata information. Users may also want to save the results in an R-specific format, for an easy return to other analyses supported by Bioconductor. In this case, consult the documentation for the [save](#) function.

Post differential analysis sanity checks. Figure 7 (Step 9) shows the typical features of a *p*-value histogram resulting from a good data set: a sharp peak at the left side, containing genes with strong differential expression, a “floor” of values that are approximately uniform in the interval $[0, 1]$, corresponding to genes that are not differentially expressed (for which

the null hypothesis is true), and a peak at the upper end, at 1, resulting from discreteness of the Negative Binomial test for genes with overall low counts. The latter component is often less pronounced, or even absent, when the likelihood ratio test is used. In addition, users should spot check genes called as differentially expressed by loading the sorted BAM files into a genome browser.

Reproducible research. So that other researchers (e.g., collaborators, reviewers) can reproduce data analyses, we recommend that users keep a record of the commands and the software versions used in their analysis. In practice, this is best achieved by keeping the complete transcript of the computer commands interweaved with the textual narrative in a single, executable document⁷⁶.

R provides many tools to facilitate the authoring of executable documents, including the [Sweave](#) function and the `knitr` package. The `sessionInfo` function helps with documenting package versions and related information. A recent integration with Rstudio is [rpubs.com](#), which provides seamless integration of “mark-down” text with R commands for easy web-based display. For language-independent authoring, a powerful tool is provided by `Emacs org-mode`.

Acknowledgements The authors wish to thank Xiaobei Zhou for comparing counting methods, Olga Nikolayeva for feedback on an earlier version of the manuscript and members of the ECCB Workshop (Basel, September 2012) for their feedback. DJM is funded by the General Sir John Monash Foundation, Australia. MDR wishes to acknowledge funding from the University of Zurich’s Research Priority Program in Systems Biology and Functional Genomics. SA, WH and MDR acknowledge funding from the European Commission through the 7th Framework Collaborative Project RADIANT (Grant Agreement Number: 305626).

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to M.D.R. or W.H. (email: mark.robinson@imls.uzh.ch, whuber@embl.de); questions concerning the use of Bioconductor software (including `edgeR` and `DESeq`) should be directed to the Bioconductor mailing list⁶⁸.

1. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628 (2008).

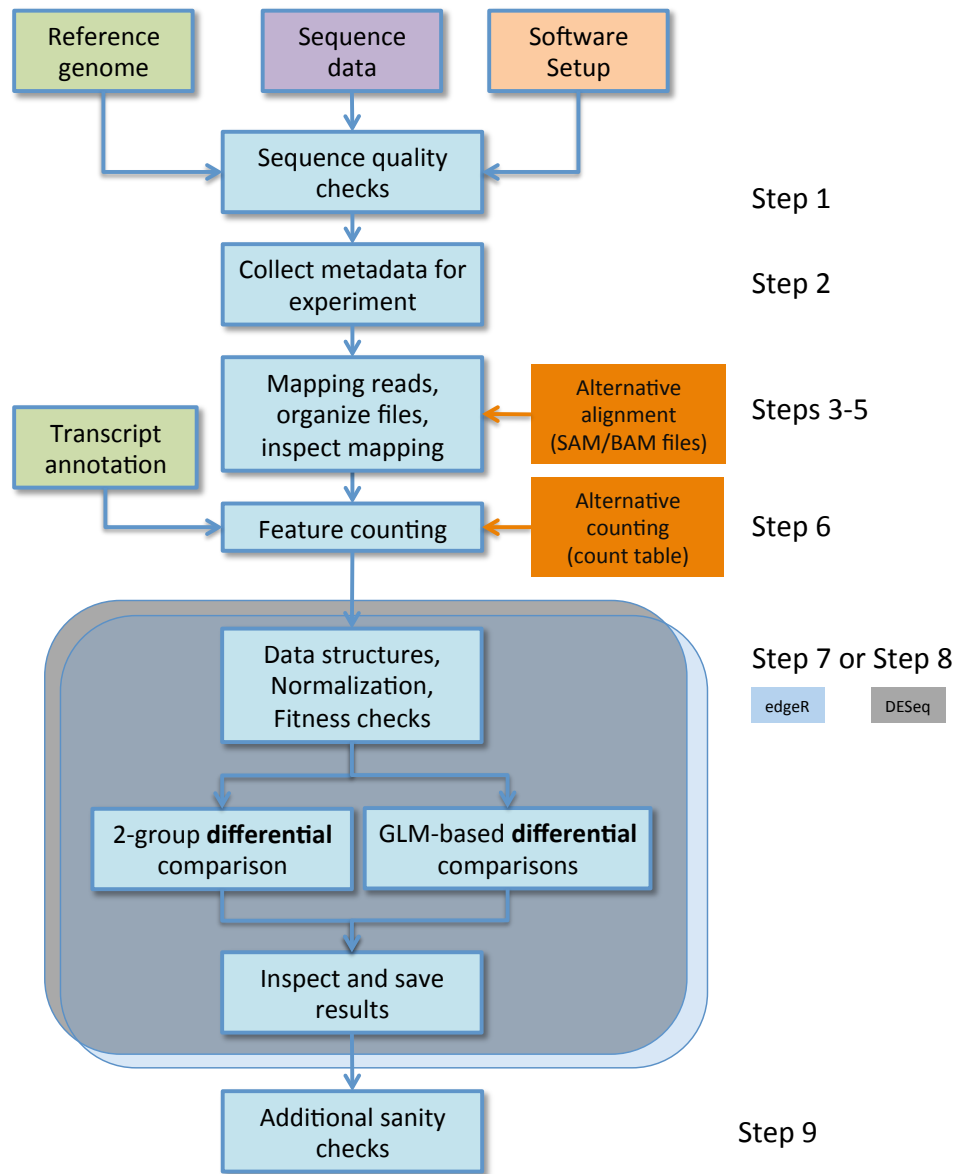


Figure 1: The pipeline for count-based differential expression analysis using **edgeR** and/or **DESeq**. Many steps are common to both tools (Steps 1-6, 9), while the specific commands are different (Step 7 for **edgeR**, Step 8 for **DESeq**). Steps within the **edgeR** or **DESeq** differential analysis can follow two paths, depending on whether the experimental design is *simple* or *complex*. Alternative entry points to the protocol are shown in orange boxes.

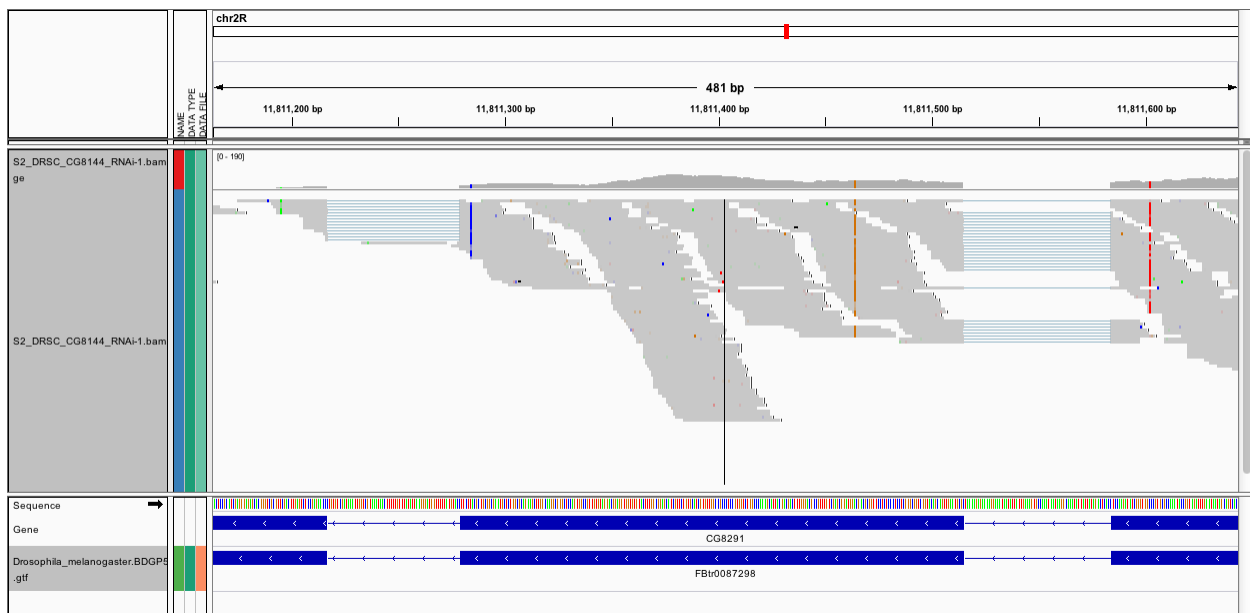


Figure 3: A screenshot of reads aligning across exon junctions.

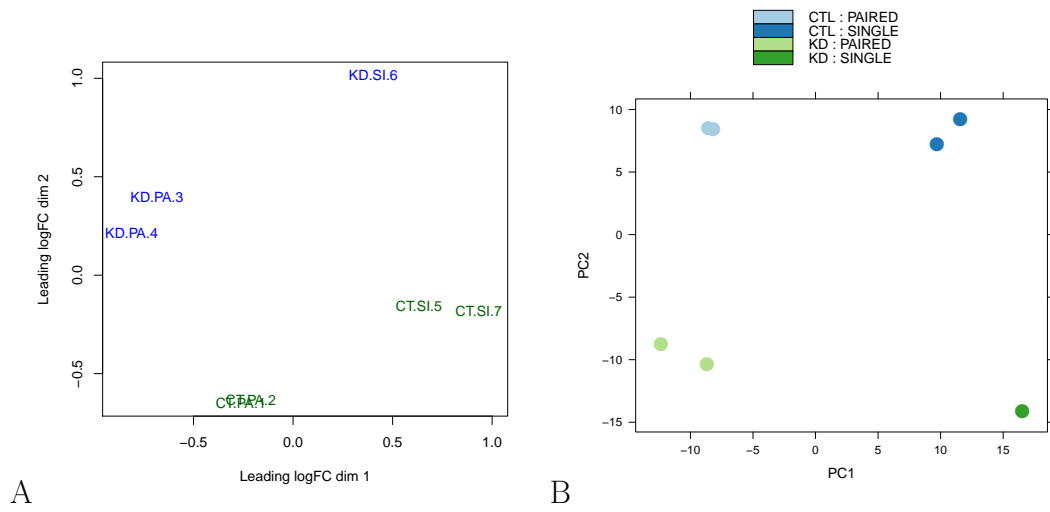


Figure 4: A. Using a count-specific distance measure, `edgeR`'s `plotMDS` produces a multi-dimensional scaling plot showing the relationship between all pairs of samples. B. `DESeq`'s `plotPCA` makes a principal component plot of vst-transformed count data.

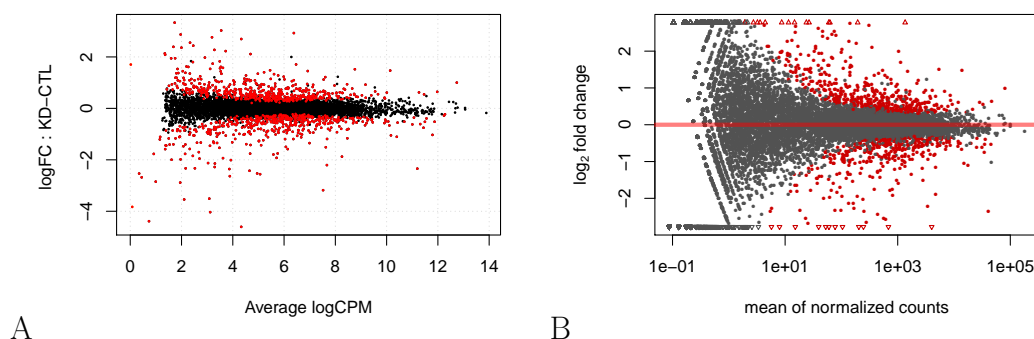


Figure 5: A. `edgeR`'s `plotSmear` function plots the log-fold change (i.e., the log ratio of normalized expression levels between two experimental conditions) against the log-counts-per-million. B. Similarly, `DESeq`'s `plotMA` displays differential expression (log-fold-changes) versus expression strength (log-average-read-count).

2. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57–63 (2009).
3. Borodina, T., Adjaye, J. & Sultan, M. *A strand-specific library preparation protocol for RNA sequencing.*, vol. 500 (Elsevier Inc., 2011).
4. Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods* **7**, 709–715 (2010).
5. Lu, C., Meyers, B. C. & Green, P. J. Construction of small RNA cDNA libraries for deep sequencing. *Methods* **43**, 110–117 (2007).
6. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11**, R106 (2010).
7. Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881–2887 (2007). URL <http://www.ncbi.nlm.nih.gov/pubmed/17881408>.
8. Robinson, M. D. & Smyth, G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9**, 321–332 (2008).

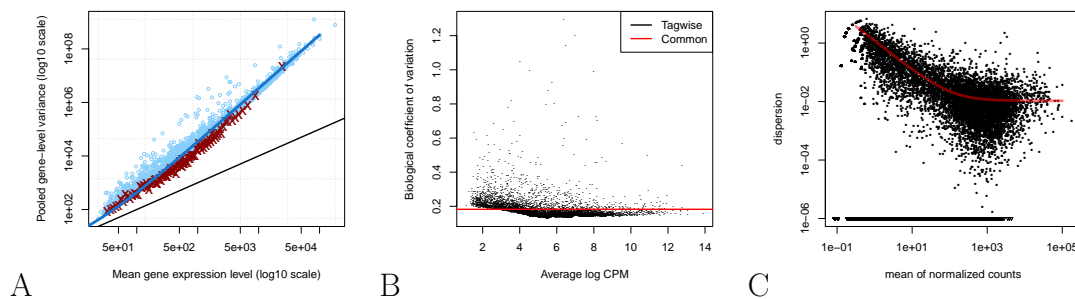


Figure 6: A. edgeR's `plotMeanVar` can be used for exploring the mean-variance relationship; each dot represents the estimated mean and variance for each gene, with binned variances as well as the trended common dispersion overlaid. B. edgeR's `plotBCV` illustrates the relationship of biological coefficient of variation versus the mean. C. DESeq's `plotDispEsts` shows the fit of dispersion versus mean.

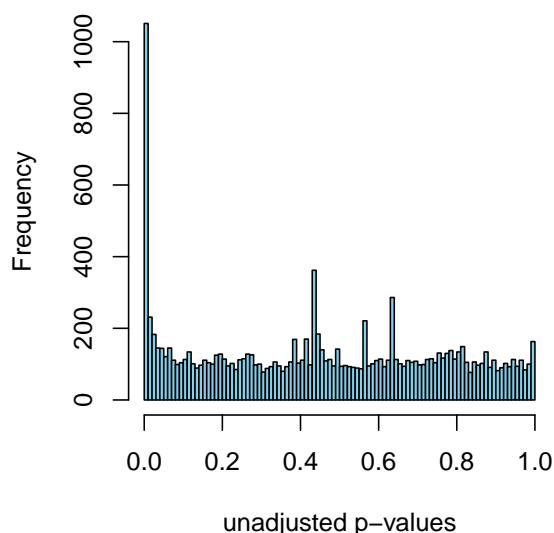


Figure 7: Histogram of p -values from gene-by-gene statistical tests.

9. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
10. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* 1–10 (2012).
11. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80 (2004).
12. Fonseca, N. A., Rung, J., Brazma, A. & Marioni, J. C. Tools for mapping high-throughput sequencing data. *Bioinformatics* (2012).
13. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**, 562–578 (2012).
14. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).

15. Van De Wiel, M. A. *et al.* Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* **14**, 113–28 (2013).
16. Wu, H., Wang, C. & Wu, Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* (2012).
17. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644–652 (2011).
18. Siebert, S. *et al.* Differential gene expression in the siphonophore *Nanomia bijuga* (Cnidaria) assessed with multiple next-generation sequencing workflows. *PLoS ONE* **6**, 12 (2011).
19. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
20. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511–515 (2010).
21. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* (2012).
22. Robinson, M. D. *et al.* Copy-number-aware differential analysis of quantitative DNA sequencing data. *Genome Research* **22**, 2489–96 (2012).
23. Vanharanta, S. *et al.* Epigenetic expansion of VHL-HIF signal output drives multiorgan metastasis in renal cancer. *Nature Medicine* **19**, 50–6 (2013).
24. Samstein, R. M. *et al.* Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell* **151**, 153–66 (2012).
25. Johnson, E. K. *et al.* Proteomic analysis reveals new cardiac-specific dystrophin-associated proteins. *PloS ONE* **7**, e43515 (2012).
26. Hardcastle, T. J. & Kelly, K. A. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**, 422 (2010).
27. Zhou, Y.-H., Xia, K. & Wright, F. a. A Powerful and Flexible Approach to the Analysis of RNA Sequence Count Data. *Bioinformatics* **27**, 2672–2678 (2011).

28. Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: A matter of depth. *Genome Research* **21**, 2213–2223 (2011).
29. Lund, S. P., Nettleton, D., McCarthy, D. J. & Smyth, G. K. Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates. *Statistical Applications in Genetics and Molecular Biology* **11**, Article 8 (2012).
30. Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C. & Brenner, S. E. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446**, 926–929 (2007).
31. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Research Advance* **Ac**, 1–19 (2012).
32. Glaus, P., Honkela, A. & Rattray, M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* **28**, 1721–8 (2012).
33. Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M. & Gilad, Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome Research* **20**, 180–189 (2010).
34. Okoniewski, M. J. *et al.* Preferred analysis methods for single genomic regions in RNA sequencing revealed by processing the shape of coverage. *Nucleic Acids Research* 1–11 (2011).
35. Illumina iGenomes. URL <http://tophat.cbcb.umd.edu/igenomes.html>.
36. Hansen, K. D., Irizarry, R. A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 204–216 (2012).
37. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics* **12**, 480 (2011).
38. Anders, S. HTSeq: Analysing high-throughput sequencing data with Python. URL <http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html>.
39. Delhomme, N., Padiou, I., Furlong, E. E. & Steinmetz, L. easyRNASeq: a bioconductor package for processing RNA-Seq data. *Bioinformatics* 2–3 (2012).
40. Aboyoun, P., Pages, H. & Lawrence, M. GenomicRanges: Representation and manipulation of genomic intervals.

41. Lerch, A., Gaiditzis, D. & Stadler, M. Quantify and Annotate Short Reads in R. URL <http://www.bioconductor.org/packages/2.12/bioc/html/QuasR.html>.
42. Hansen, K. D., Wu, Z., Irizarry, R. a. & Leek, J. T. Sequencing technology does not eliminate biological variability. *Nature Biotechnology* **29**, 572–573 (2011).
43. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11**, 733–739 (2010).
44. Auer, P. L. & Doerge, R. W. Statistical design and analysis of RNA sequencing data. *Genetics* **185**, 405–416 (2010).
45. Gagnon-Bartsch, J. A. & Speed, T. P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539–52 (2011).
46. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics* **3**, 12 (2007).
47. Smyth, G. K. Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, 397–420 (Springer, New York, 2005).
48. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**, R25 (2010).
49. Nookaew, I. *et al.* A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. *Nucleic Acids Research* **40**, 10084–97 (2012).
50. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **in press** (2013).
51. Rapaport, F. *et al.* Comprehensive evaluation of differential expression analysis methods for RNA-seq data. *arXiv* 1301.5277v2 (2013).
52. Cygwin. URL <http://www.cygwin.com/>.
53. Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotechnology* **27**, 455–457 (2009).
54. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).

55. Wang, K. *et al.* MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research* **38**, e178 (2010).
56. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research* **41** (2013). URL <http://subread.sourceforge.net>.
57. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **Advance pu**, bbs017– (2012).
58. Fiume, M., Williams, V., Brook, A. & Brudno, M. Savant: genome browser for high-throughput sequencing data. *Bioinformatics* **26**, 1938–44 (2010).
59. Fiume, M. *et al.* Savant Genome Browser 2: visualization and analysis for population-scale genomics. *Nucleic Acids Research* **40**, 1–7 (2012).
60. R Development Core Team, R. R: A Language and Environment for Statistical Computing (2011). URL <http://www.r-project.org>.
61. Morgan, M. *et al.* ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**, 2607–2608 (2009).
62. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
63. Brooks, A. N. *et al.* Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Research* **21**, 193–202 (2011).
64. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**, 207–210 (2002).
65. Cox, D. R. & Reid, N. Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistical Society Series B Methodological* **49**, 1–39 (1987).
66. Dudoit, S., Yang, Y. H., Callow, M. J. & Speed, T. P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–139 (2002).

67. Leisch, F. Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis. In Härdle, W. & Rönz, B. (eds.) *Compstat 2002 Proceedings in Computational Statistics*, no. 69 in *Compstat 2002 - Proceedings in Computational Statistics*, 575–580. Institut für Statistik und Wahrscheinlichkeitstheorie, Technische Universität Wien (Physica Verlag, Heidelberg, 2002).
68. Bioconductor mailing list.
69. Bioconductor mailing list Posting Guide. URL <http://bioconductor.org/help/mailling-list/posting-guide/>.
70. FASTX-Toolkit. URL http://hannonlab.cshl.edu/fastx_toolkit/.
71. Andrews, S. Fastqc: A quality control tool for high throughput sequence data. URL <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
72. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 9546–9551 (2010).
73. Cappiello, C., Francalanci, C. & Pernici, B. Data quality assessment from the user’s perspective. *Architecture* **22**, 68–73 (2004).
74. Myers, R. M. *Classical and Modern Regression with Applications* (Duxbury Classic Series, 2000), 2nd editio edn.
75. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological* **57**, 289–300 (1995).
76. Gentleman, R. Reproducible research: a bioinformatics case study. *Statistical applications in genetics and molecular biology* **4**, Article2 (2005).